ANÁLISE DE CORRESPONDÊNCIA APLICADA À PESQUISA EM ENSINO DE CIÊNCIAS

Matheus Monteiro Nascimento, Cláudio Cavalcanti, Fernanda Ostermann
Universidade Federal do Rio Grande do Sul
matheus.monteiro@ufrgs.br, claudio.cavalcanti@ufrgs.br, fernanda.ostermann@ufrgs.br

RESUMO: A área de Ensino de Ciências no Brasil é um campo de pesquisa que se vale de múltiplas formas de coletas de dados, como por exemplo, entrevistas, questionários, testes padronizados, entre outras. São variadas também as maneiras de analisar e interpretar esses dados. Neste trabalho nos propomos a discutir a utilização de uma ferramenta para análise estatística de dados denominada Análise de Correspondência. Para exemplificar a sua aplicação no campo educacional utilizamos os microdados do ENEM de 2014, pois apresentam uma expressiva quantidade de variáveis que não podem ser relacionadas sem um método estatístico adequado de análise multivariada. Os resultados obtidos com a Análise de Correspondência na interpretação dos microdados corroboram resultados encontrados na literatura recente e permite vislumbrar novos aspectos importantes, sugerindo a validade da análise realizada.

PALAVRAS CHAVE: Análise de Correspondência; estatística; microdados; ENEM.

OBJETIVOS: Neste trabalho objetivamos apresentar um método para análise estatística multivariada chamada Análise de Correspondência (AC), pouco utilizado na área de Ensino de Ciências, articulando-o ao referencial teórico boudesiano para em conjunto com esse referencial explicar o perfil de candidatos bem sucedidos no ENEM de 2014. A Análise de Correspondência já é utilizada em outras áreas do conhecimento desde meados do século XX, contudo, queremos discutir neste texto de que maneira ela pode ser empregada e articulada a uma fundamentação teórica na exploração de possíveis associações entre diversas variáveis em uma pesquisa do Ensino de Ciências. A AC é um método destinado especificamente para a análise de associação entre variáveis categóricas. Variáveis numéricas podem ser convertidas em categóricas por discretização, via atribuição de categorías ordinais que correspondem a intervalos determinados das escalas de valores dessas variáveis (podem ser contínuas ou inteiras). Esses valores escalonados apresentam propriedades geométricas interessantes e permitem a construção do chamado mapa de relação entre as variáveis. Na sequência do texto apresentaremos a ferramenta de análise em maior detalhe e, em seguida, ilustraremos sua aplicação com um exemplo utilizando os microdados do ENEM de 2014.

ANÁLISE DE CORRESPONDÊNCIA (AC)

A Análise de Correspondência Simples (AC) é uma técnica multivariada para análise exploratória de dados categorizados em tabelas de contingência de duas vias (duas variáveis categóricas), levando em

conta algumas medidas de correspondência entre a variável cujas categorias são descrita nas linhas e a variável cujas categorias são descrita nas colunas (Greenacre e Blasius, 2006). Sua origem matemática aparece no trabalho de Hirschfield (1935) e, desde então, seus procedimentos numéricos e algébricos têm sido utilizados em diferentes contextos, como na ecologia e psicologia. Por permitir a identificação de múltiplos fatores que são pertinentes ao fenômeno de estudo e por ser uma técnica descritiva e apropriada à análise de variáveis categóricas, a AC é uma importante ferramenta analítica também para as ciências sociais (Greenacre e Blasius, 1994), como podemos observar nos trabalhos de Bourdieu (1998), Greenacre e Blasius (1994) e Ferreira (2003). A AC permite investigar relações entre as duas variáveis categóricas de uma tabela de contingência, por meio da associação entre as respectivas categorias. Ela possui diversas características que a distinguem de outras técnicas de análise. Uma dessas características é a sua natureza multivariada, que permite investigar relações que não são facilmente percebidas a partir da simples comparação de pares de variáveis. A única exigência da AC é que os dados sejam todos positivos e dispostos em uma tabela retangular. Quando a tabela de contingência possui duas variáveis de entrada utilizamos a AC simples. Quando mais variáveis categóricas se fazem presentes, utilizamos a sua forma mais geral, chamada Análise de Correspondência Múltipla (ACM).

São encontrados na literatura especializada diferentes métodos para se realizar a ACM, por exemplo, o método da matriz indicadora Z e da matriz de *Burt*. Os resultados obtidos utilizando esses dois métodos são relativamente próximos. Em nosso trabalho utilizamos uma ACM um pouco diferente, que é a chamada Análise de Correspondência Múltipla Conjunta (ACMC), ou simplesmente Análise de Correspondência Conjunta (ACC). Camiz e Gomes (2013) discutem as diferenças na utilização da ACM padrão e da ACC e concluem, a partir dos resultados de duas aplicações, que a ACC é a melhor técnica a ser utilizada. Além disso, Greenacre (1988) desenvolveu a ACC tendo em vista que na ACM certas operações resultavam valores superestimados para a inercia (variância) total, fazendo com que as contribuições das dimensões principais fossem subestimadas.

O procedimento para interpretar as dimensões obtidas com a ACM não é trivial e foge do escopo desse trabalho. Greenacre (1991) e Knop (2008) apresentam alguns exemplos de como realizar essa interpretação. Para os objetivos desse trabalho não é fundamental significar as dimensões da ACM, mas sim, investigar as distâncias relativas entre as variáveis e suas posições no mapa gerado.

UM EXEMPLO DE APLICAÇÃO

Em síntese, a ACC é usualmente indicada para representar matrizes com múltiplos dados categóricos e sem uma estrutura claramente definida. Este método permite que se visualizem as associações mais importantes entre múltiplas de variáveis. Os resultados podem ser apresentados em gráficos, onde se representam as categorias de cada variável e onde podem ser observadas as associações entre elas, através da distância entre os pontos desenhados (Greenacre, 1981; Lebart *et al.*, 1984). Neste trabalho, vamos exemplificar a aplicação da ACC analisando os dados do ENEM. Todos os anos o INEP disponibiliza arquivos com os microdados do ENEM realizado no ano anterior, entre os quais são disponibilizadas as respostas dadas pelos alunos a um questionário socioeconômico preenchido no ato da inscrição do exame. As mais de 180 perguntas do questionário estão relacionadas com o contexto familiar e escolar dos alunos. Vamos entender a importância dessas variáveis de contexto a partir da aplicação da ACC. Em nosso exemplo, as linhas da matriz são os alunos que realizaram ENEM em 2014 - que correspondem aos últimos dados fornecidos pelo INEP - e as colunas são variáveis de contexto relacionadas com os estudantes. A tabela 1 mostra 35 das mais de 440000 linhas e as 5 colunas que constituem os dados utilizados na ACM. Nos microdados de 2014 há originalmente mais de 8 milhões de candidatos. Foram filtrados apenas os que preencheram completamente o questionário socioeconômico, que declararam sua etnia, que estavam viculados a uma escola e que tivessem comparecido a todas as provas objetivas.

Tabela 1.
Fragmento dos dados utilizados na Análise
de Correspondência Múltipla. Fonte: INEP/2014

NU_INSCRICAO	DEP_ADM	ETNIA	DES_ENEM	ICE_ICC
140000000063	Estadual	Pardo(a)	Fraco [439.9 - 484.2)	CE4_CC2
140000000071	Estadual	Branco(a)	Bom [530.7 - 590.9)	CE5_CC6
140000000079	Estadual	Pardo(a)	Fraco [439.9 - 484.2)	CE6_CC4
140000000120	Municipal	Negro(a)	Fraco [439.9 - 484.2)	CE6_CC4
140000000288	Estadual	Branco(a)	Regular [484.2 - 530.6)	CE1_CC1
140000000440	Estadual	Pardo(a)	Regular [484.2 - 530.6)	CE3_CC5
140000000441	Estadual	Branco(a)	Bom [530.7 - 590.9)	CE6_CC4
140000000466	Estadual	Pardo(a)	Regular [484.2 - 530.6)	CE7_CC7
140000000496	Estadual	Pardo(a)	Fraco [439.9 - 484.2)	CE4_CC2
140000000539	Estadual	Pardo(a)	Bom [530.7 - 590.9)	CE2_CC3
140000000541	Estadual	Negro(a)	Fraco [439.9 - 484.2)	CE2_CC3
140000000555	Privada	Branco(a)	Bom [530.7 - 590.9)	CE7_CC7
140000000558	Estadual	Pardo(a)	Bom [530.7 - 590.9)	CE4_CC2
140000000561	Municipal	Branco(a)	Ótimo [590.9 - 823.7)	CE5_CC6
140000000586	Municipal	Pardo(a)	Regular [484.2 - 530.6)	CE3_CC5
140000000610	Estadual	Pardo(a)	Fraco [439.9 - 484.2)	CE4_CC2
140000000659	Estadual	Branco(a)	Regular [484.2 - 530.6)	CE3_CC5
140000000670	Estadual	Branco(a)	Bom [530.7 - 590.9)	CE3_CC5
140000000714	Municipal	Branco(a)	Fraco [439.9 - 484.2)	CE4_CC2
140000000768	Estadual	Pardo(a)	Regular [484.2 - 530.6)	CE3_CC5
140000000793	Estadual	Pardo(a)	Fraco [439.9 - 484.2)	CE5_CC6
140000000853	Estadual	Negro(a)	Fraco [439.9 - 484.2)	CE2_CC3
140000000861	Estadual	Pardo(a)	Ótimo [590.9 - 823.7)	CE5_CC6
140000000869	Federal	Negro(a)	Regular [484.2 - 530.6)	CE5_CC6
140000000874	Estadual	Negro(a)	Ótimo [590.9 - 823.7)	CE3_CC5
140000000924	Estadual	Pardo(a)	Regular [484.2 - 530.6)	CE5_CC6
140000001013	Estadual	Pardo(a)	Fraco [439.9 - 484.2)	CE1_CC1
140000001071	Estadual	Branco(a)	Ruim [324.2 - 439.9)	CE6_CC4
140000001075	Estadual	Pardo(a)	Regular [484.2 - 530.6)	CE3_CC5
140000001077	Municipal	Branco(a)	Ruim [324.2 - 439.9)	CE2_CC3
140000001104	Estadual	Pardo(a)	Fraco [439.9 - 484.2)	CE4_CC2
140000001131	Estadual	Pardo(a)	Regular [484.2 - 530.6)	CE5_CC6
140000001133	Estadual	Pardo(a)	Fraco [439.9 - 484.2)	CE2_CC3
140000001181	Estadual	Branco(a)	Regular [484.2 - 530.6)	CE7_CC7
140000001215	Estadual	Branco(a)	Bom [530.7 - 590.9)	CE6_CC4

As variáveis que utilizamos para a ACC e que aparecem na tabela 1 foram o número de inscrição que identifica o candidato (NU_INSCRICAO), a dependência administrativa da escola em que o estudante concluiu o Ensino Médio (DEP_ADM), a etnia autodeclarada pelo candidato (ETNIA), o desempenho obtido no exame (DES_ENEM) e uma variável que denominamos índice de capital econômico e cultural (ICE_ICC), que é constituída por itens relacionados com o capital econômico – como renda familiar média e bens materiais – e com o capital cultural – como nível de instrução dos pais e outros itens, todos relacionados ao capital cultural institucionalizado (Bourdieu, 1986, p. 47). Estes questionários foram analisados com o modelo de respostas graduadas da Teoria da Resposta ao Item (Samejima, 1969), investigando-se sua estrutura fatorial (dimensionalidade) e calculando-se ao final um escore numérico para o capital econômico e cultural de cada candidato. Para discretizar essas duas variáveis e criar a variável ICE_ICC, fizemos uma análise de cluster *k-means* (Izenman, 2008) nos escores da capital económico e cultural, classificando os índices de capital econômico e cultural em 7 grupos diferentes. A divisão segue a seguinte classificação: muito baixo (1), baixo (2), médio-baixo (3), médio (4), médio-alto (5), alto (6) e muito-alto (7). A figura 1 ilustra essa divisão graficamente. Com isso, o grupo "pior" classificado é o que possui capital econômico 1 e capital cultural também 1 (CE1_CC1). O grupo "melhor" classificado é o que possui capital econômico 7 e capital cultural também igual a 7 (CE7_CC7). Os outros grupos assumem níveis diferentes de capitais, tanto econômicos, quanto culturais. A variável DES_ENEM foi classificada também por meio de análise de cluster *k-means* na média geral do candidato (média dos escores nas provas objetivas), com 5 níveis: ruim, fraco, regular, bom e ótimo.

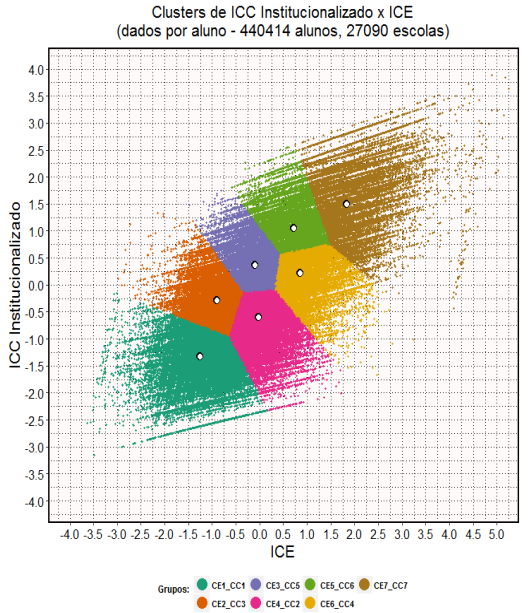

Fig. 1. Cluster hierárquico para a associação dos índices de capital econômico e cultural (ICE_ICC). *Fonte*: INEP/2014.

Com base no trabalho de Nenadic e Greenacre (2007) fizemos a ACC a partir dos dados da tabela 1. O gráfico da figura 2 permite inferir a associação entre as variáveis (categorias da coluna na tabela 1) tomando todos os candidatos (linhas) considerados. A análise revela, de maneira contundente, as desigualdades sociais do sistema educacional brasileiro. Notamos que candidatos pertencentes a grupos étnicos como negros e pardos, oriundos de escolas com dependência administrativa estadual e com baixas categorias de capital econômico e cultural, estão fortemente relacionados com desempenhos Fraco e Ruim no ENEM. Setas cujo comprimento é pequeno (descrevendo pontos próximo da origem) não contribuem significativamente para a informação contida no gráfico (que explica 94,81 por cento da inércia – ou variância – total). Esse é o caso das categorias de etnia Amarela e dependência administrativa municipal, por exemplo. A orientação relativa entre as setas, que representam cada categoria, permite inferir sobre o grau de associação. Por exemplo, candidatos oriundos de escolas federais estão bem mais fortemente associados a desempenho Ótimo no ENEM do que escolas estaduais, que estão entre o desempenho Fraco e Regular. Como esperado, candidatos da categoria mais alta de capital econômico e cultural (CE7_CC7) estão mais associados a desempenho bom e ótimo do que os demais, ao passo que candidatos do mais baixa (CE1_CC1) estão mais associados a desempenho ruim. Esse resultado mostra que a adoção do ENEM como sistema de ingresso ao ensino superior reforça as desigualdades sociais existentes, mantendo o acesso às universidades federais muito mais ao alcance de candidatos de etnia branca com elevado capital econômico e cultural, majoritariamente oriundos de escolas privadas e federais. Essa conclusão não poderia ser obtida simplesmente olhando para os dados da tabela 1. O gráfico obtido a partir da ACC revelou um cenário até certo ponto esperado, uma forte relação entre variáveis de contexto e o desempenho em exames de larga escala, como apontado por Alves e Soares (2007), Andrade e Laros (2007), Klein *et al.* (2007), Golgher (2010), Travitzki (2013) e Silveira, Barbosa e da Silva (2015). No entanto, esse estudo vai além dos citados na medida em que estuda associação simultânea entre mais categorias, fundamentado no poder da ACC para análise multivariada de dados. A concordância entre os resultados da ACC com os trabalhos encontrados na literatura é um

indicador de que essa ferramenta estatística pode, de fato, ser uma aliada dos pesquisadores em Ensino de Ciências, na medida em que muitas pesquisas da área envolvem questionários semelhantes.

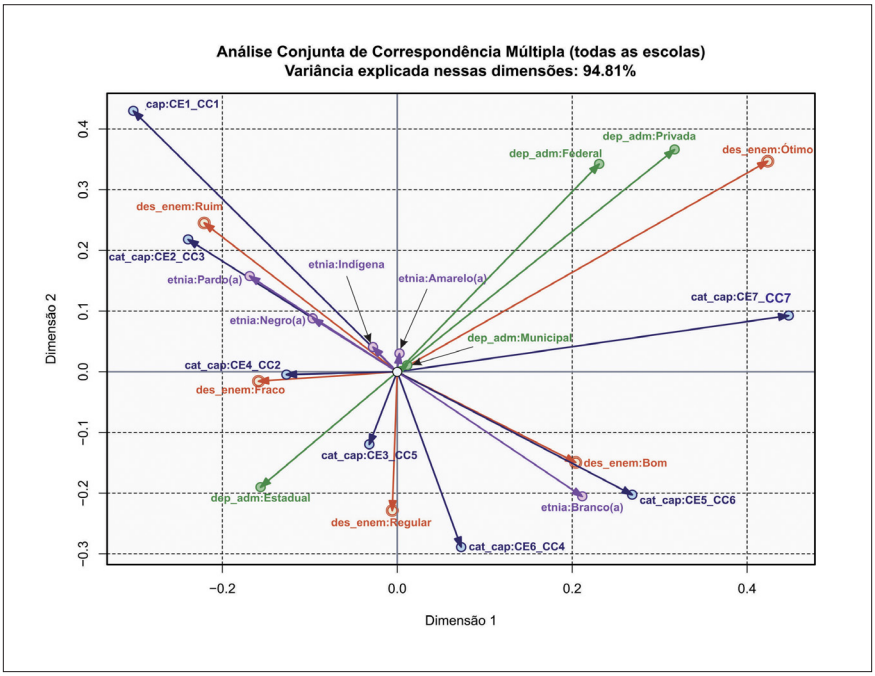

Fig. 2. Análise de Correspondência Múltipla entre as variáveis desempenho no ENEM, etnia, dependência administrativa e categorias de capital econômico/ cultural.

CONCLUSÕES

A área de Ensino de Ciências no Brasil se constituiu a partir da intersecção das Ciências da Natureza – Biologia, Física e Química – com a Educação e a Psicologia, principalmente. Discutimos nesse trabalho a articulação da Análise de Correspondência Conjunta (ACC) ao referencial teórico bourdesiano na investigação de evidências para a associação entre capitais econômico e cultural institucionalizado no desempenho no ENEM 2014, obtidas dos microdados. A ACC, por permitir investigar associações entre múltiplas variáveis categóricas, permite traçar um perfil de quem é bem ou mal sucedido em um exame de grande porte como o ENEM e evidenciar desigualdades.

As variáveis consideradas aqui foram além dos índices de capitais econômico e cultural institucionalizado de Bourdieu, embora relacionadas com eles: a dependência administrativa da escola em que o estudante concluiu o Ensino Médio, a etnia autodeclarada pelo candidato, o desempenho obtido no exame. A análise revelou uma grave desigualdade social no sistema educacional brasileiro, como apontado em outros estudos, indicando que a ACC realizada levou a resultados consistentes e, além disso, permitido que mais variáveis fossem levadas em conta simultaneamente.

Propomos nesse texto a utilização de uma ferramenta para análise estatística de dados. que exige o olhar teórico do pesquisador na hora de interpretar as relações que são obtidas. Ou seja, é fundamental um referencial teórico que auxilie na interpretação dos resultados estatísticos.

REFERÊNCIAS

Alves, M. T. G.; Soares, J. F. As pesquisas sobre o efeito das escolas: contribuições metodológicas para a sociologia da educação. Sociedade e Estado, Brasília, v. 22, n. 2, p. 435-473, 2007.

Bourdieu, P. The forms of capital. In: Richardson, J. (Ed.). Handbook of theory and research for the sociology of education. New York: Greenwood, 1986. p. 241-258.

Bourdieu, P. The state nobility: Elite schools in the field of power. Stanford University Press, 1998.

Camiz, S.; Gomes, G. C. Joint Correspondence Analysis versus Multiple Correspondence Analysis: a solution to an undetected problem. In: Giusti, A.; Ritter, G.; Vichi, M. (Eds.). Classification and data mining. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 11-18.

de Andrade, J. M.; Laros, J. A. Fatores associados ao desempenho escolar: estudo multinível com dados do SAEB/2001. Psicologia: Teoria e Pesquisa, v. 23, n. 1, p. 033-042, 2007.

Ferreira, M. C. Consumo cultural e espaços sociais: os vestibulandos das universidades públicas na cidade do Rio de Janeiro, 1990. Opinião Pública, v. 9, n. 1, p. 170-189, 2003.

Greenacre, M. Correspondence analysis of multivariate categorical data by weighted least-squares. Biometrika, v. 75, n. 3, p. 457-467, 1988.

— Interpreting multiple correspondence analysis. Applied Stochastic Models and Data Analysis, v. 7, n. 2, p. 195-210, 1991.

Greenacre, M.; Blasius J. *et al*. Correspondence analysis in the social sciences: Recent developments and applications. 1994.

— (Ed.). Multiple correspondence analysis and related methods. CRC press, 2006.

Golgher, A. Diálogos com o ensino médio 6: o estudante de ensino médio no Brasil analisado a partir de dados do INEP. Belo Horizonte: UFMG/Cedeplar, 2010.

Izenman, A. J. Modern multivariate statistical techniques: regression, classification, and manifold learning. New York: Springer, 2008.

Klein, R.; Fontanive, N. S.; Elliot, L. G. O exame nacional do ensino médio–Tecnologia e principais resultados em 2005. REICE: Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 2007.

Knop, M. N. A escolha de curso superior dos vestibulandos da Universidade Federal do Rio Grande do Sul: um estudo quantitativo com utilização de análise de correspondência múltipla. Tese de Doutorado. Universidade Federal do Rio Grande do Sul. 2008.

Lebart, L. Complementary use of correspondence analysis and cluster analysis. Correspondence analysis in the social sciences, p. 162-178, 1994.

Nenadic, O.; Greenacre, M. Correspondence analysis in R, with two-and three-dimensional graphics: The ca package. Journal of Statistical Software, v. 20, n. 3. 2007.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, v. 34, n. 1, p. 1-97, 1969.

Silveira, F. L. da; Barbosa, M. C. B.; Silva, R. da. Exame Nacional do Ensino Médio (ENEM): Uma análise crítica. Revista Brasileira de Ensino de Física, v. 37, n. 1, p. 1101, 2015.

Travitzki, R.. ENEM: limites e possibilidades do Exame Nacional do Ensino Médio enquanto indicador de qualidade escolar. 322 f. 2013. Tese (Doutorado em Educação)–Faculdade de Educação, Universidade de São Paulo, São Paulo.